

X-cDiT: Explainable Latent Conditional Diffusion Inpainting with Transformers for Iris Occlusion Removal

Negoîtescu Andreea, Czibula Gabriela, Mircea Cristina
Babeş-Bolyai University

WeADL 2026 Workshop

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number [PN-IV-P7-7.1-PED-2024-0121](#), within PNCDI IV



Contents

- 1 Background and Motivation
- 2 Proposed Approach
- 3 Objectives
- 4 Original Contributions
- 5 Related Work
- 6 Methodology
- 7 Results
- 8 Comparison to Related Work
- 9 Explainability
- 10 Practical Utility
- 11 Conclusions
- 12 Future Work

Background and Motivation

- Rising demand for secure and reliable biometric identification systems.
- Biometric data quality is a critical aspect.
- Among all biometric features, the iris is superior due to its complexity, stability and uniqueness.
- **Problem:** iris images are affected by occlusions:
 - reduce visibility of essential texture regions
 - diminish available discriminative iris features
 - add artifacts with corrupted information
 - affect iris recognition systems performance
- **Solution:** generating, through inpainting, iris patterns that fill in the occluded areas while explicitly preserving identity.

X-cDiT: Explainable Latent Conditional Diffusion Inpainting Model with Transformers:

- **Diffusion with Transformers:** training stability, contextual awareness, high visual fidelity.
- **Conditioning:** inpainting guided by the iris identity class embedding.
- **Latent Space Compression:** increased efficiency.
- **Explainability:** analysis of inpainting errors from computational and biological perspectives.

Objectives

- ① To successfully build the X-cDiT iris inpainting model.
- ② To provide answers to the research questions:
 - **RQ1.** Can an approach using latent conditional diffusion inpainting with Transformers uncover relevant patterns for restoring occluded iris regions from both near-infrared (NIR) and visible light spectrum (VIS) images?
 - **RQ2.** Does the approach introduced to answer RQ1 significantly outperform related approaches for restoring occluded iris regions?
 - **RQ3.** Can we obtain additional insights about the iris image inpainting process by analyzing the inpainting scores and errors, and how consistent are these interpretations with a biological perspective?
- ③ To develop the EYERIS software application that integrates the proposed inpainting model.

Original Contributions

- Using a conditional diffusion model guided by identity embeddings for iris occlusion inpainting.
- Integrating a Transformer architecture in the diffusion model for this exact purpose.
- Using images of 512×512 pixels and the I-SOCIAL-DB dataset.
- Outperforming existing studies on VIS and NIR.
- Assessing explainability in collaboration with the Department of General and Population Genetics within the Faculty of Biology and Geography from Babeş-Bolyai University.

- Existing work on occluded iris region reconstruction is limited and most studies do not analyze the influence of both NIR and VIS spectrums on it + resolutions up to 256×256 are used.
- Only existing study for iris diffusion inpainting is "I3FDM - IRIS Inpainting via Inverse Fusion of Diffusion Models" [LZ⁺24] (no identity conditioning or Transformers).
- Artificially generated square-shaped occlusions
 - [ZCGZ21]: Iris texture repair framework based on PGGAN (Progressive Growing GANs)
 - [CZX⁺23a]: Hybrid A-GAN (Hybrid Attention GAN)
- Nonuniform synthetic missing iris regions
 - [CZX⁺23b]: TSDRA GAN (Two-Stage Deep Residual GAN)
 - [CXC⁺24] TT-GAN (Two-Stage and Two-Discriminator GAN) NIR+VIS
 - [DCC⁺25] HLFG (High-Low Feature Fusion GAN)

Methodology - Datasets Preprocessing

- VIS: I-SOCIAL-DB [L⁺21] (400 identities)
- NIR: CASIA-IrisV4-Interval [CAS10] (249 identities)

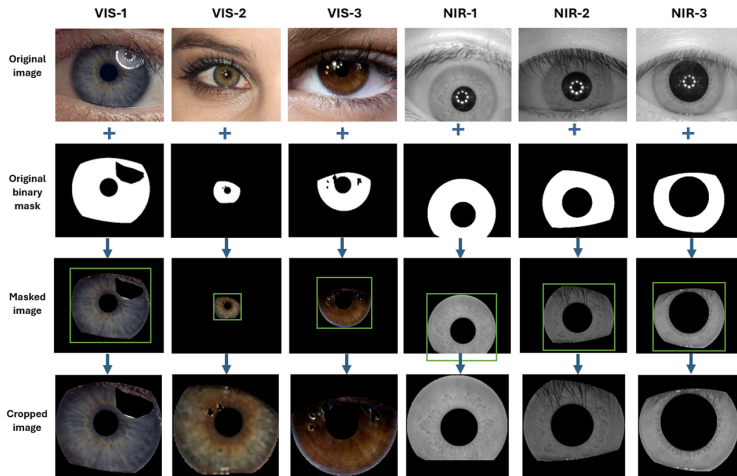


Figure: Preprocessing steps for VIS and NIR images.

Methodology - Identity Embeddings Generation I

- Encoder based on ResNet-50 [H⁺16] pretrained on ImageNet:
ResNet50 (2048-d) → Linear(2048, 1024) → ReLU → Linear(1024, 512) → L2-normalize
- Stratified split per identity 8:2
- Augmentations: ($\pm 30^\circ$, $\pm 60^\circ$, $\pm 90^\circ$)
- Training: triplet loss with margin 0.3 and contrastive learning.
- Computed mean of embeddings per identity.
- Cosine similarities:
 - VIS (-0.7868 to 0.9944, mean=0.0184, std=0.2905)
 - NIR (-0.7304 to 0.9708, mean=0.0123, std=0.2792)

Methodology - Identity Embeddings Generation II

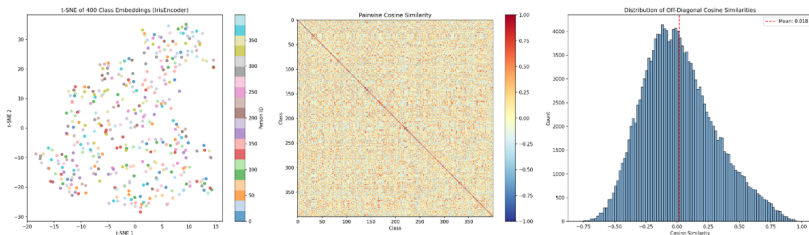


Figure: Distribution of iris identity embeddings for VIS images.

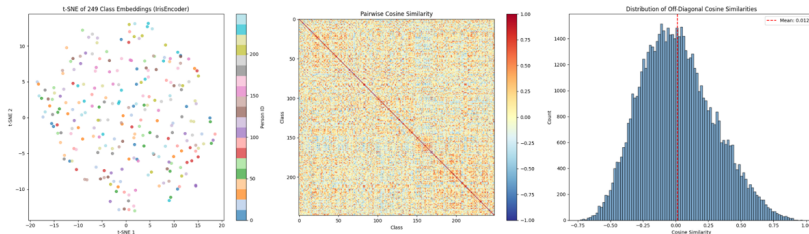


Figure: Distribution of iris identity embeddings for NIR images.

Methodology - Synthetic Occlusions Generation

VIS: 9 versions per preprocessed image $\rightarrow$ 29574 samples

NIR: 12 versions per preprocessed image $\rightarrow$ 31632 samples

Balanced generation: Small(1-2) + Medium(2-3) + Large(3-4)

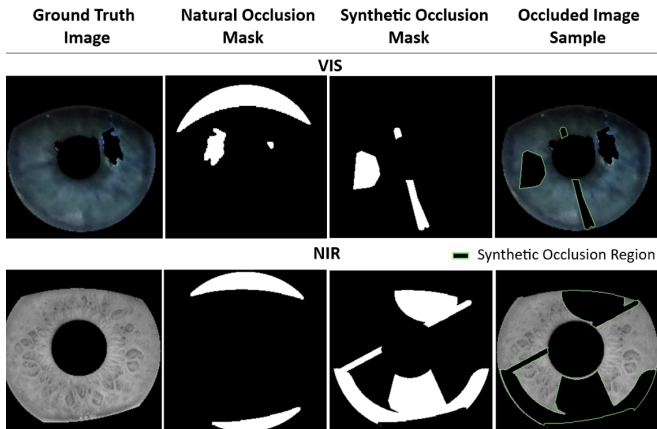


Figure: Examples of synthetic occlusions for VIS and NIR images.

Methodology - Latent Space Compression

- X-cDiT operates in latent space by performing, during generation, iterative denoising steps to progressively reconstruct the occluded iris regions.
- Pre-trained Variational Autoencoder (VAE) adapted for Stable Diffusion and fine-tuned via MSE (ft-MSE) [[Sta22](#)].
- Ground truths and occluded samples compressed:
 $512 \times 512 \times 3 \rightarrow 64 \times 64 \times 4$
- Masks downsampled to the latent resolution:
 $512 \times 512 \times 1 \rightarrow 64 \times 64 \times 1$

Methodology - The Proposed X-cDiT model

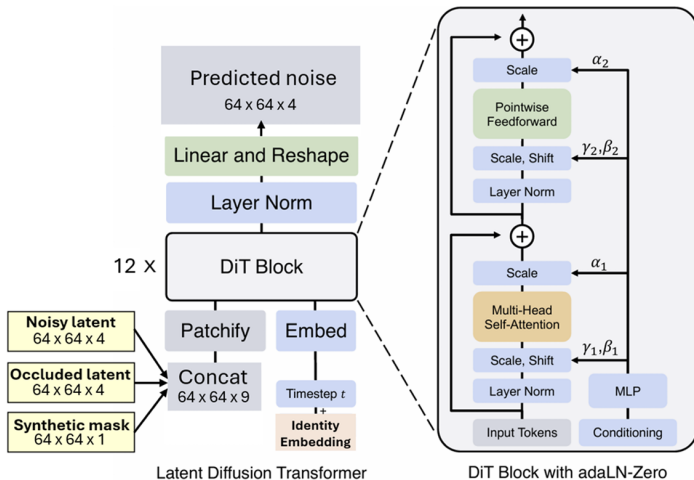


Figure: X-cDiT model architecture based on DiT B/2 [PX23].

- Latent denoising: 150 DDIM steps instead of 1000
- Classifier-free guidance (CFG) scale = 1.5

$$\epsilon_{predicted} = \epsilon_{uncond} + 1.5 \cdot (\epsilon_{cond} - \epsilon_{uncond}) \quad (1)$$

- Soft transition (expansion 3 pixels + Gaussian blur radius 2)

$$i_{final} = i_{decoded} \times m_{soft} + i_{occluded} \times (1 - m_{soft}) \quad (2)$$

Methodology - Training and Testing

- Fixed train:validation split of 8:2, stratified by identity (all occluded variants of a ground truth are in the same split)
- NVIDIA A100-SXM4-80GB GPU
- 90 epochs, batch size of 64
- AdamW optimization
- Linear warmup on the first 2000 training steps (lr 0 to $1e-4$)
- Cosine annealing scheduling (lr $1e-4$ to 0)
- Weight decay of 0.01
- Gradient clipping with norm 1.0

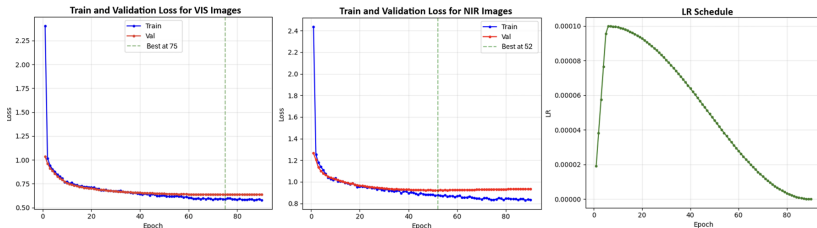


Figure: Losses and learning rate progress throughout epochs.

Results I

- Better performance for smaller mask ratios and VIS dataset
- VIS images provide more learnable signal
- VAE ft-MSE originally fine-tuned on natural RGB images

Dataset	Mask	Metrics per full image including the inpainted area				Metrics per inpainted area exclusively	
		LPIPS↓	SSIM↑	PSNR↑	MAE↓	PSNR _{masked} ↑	MAE _{masked} ↓
VIS	0-5%	0.010 ± 0.007	0.994 ± 0.005	45.09 ± 5.25	0.001 ± 0.001	28.82 ± 4.35	0.030 ± 0.013
	5-10%	0.028 ± 0.009	0.981 ± 0.009	37.14 ± 3.01	0.004 ± 0.001	26.10 ± 2.97	0.039 ± 0.014
	10-20%	0.054 ± 0.016	0.961 ± 0.017	32.73 ± 2.75	0.008 ± 0.003	24.59 ± 2.63	0.046 ± 0.015
	20-35%	0.082 ± 0.020	0.941 ± 0.025	30.04 ± 2.39	0.013 ± 0.004	23.74 ± 2.40	0.050 ± 0.016
	ALL	0.026 ± 0.023	0.982 ± 0.018	40.24 ± 7.03	0.004 ± 0.004	27.16 ± 4.17	0.036 ± 0.016
NIR	0-5%	0.011 ± 0.007	0.985 ± 0.009	41.20 ± 3.72	0.002 ± 0.001	26.32 ± 3.06	0.038 ± 0.014
	5-10%	0.030 ± 0.009	0.959 ± 0.012	35.12 ± 2.16	0.005 ± 0.001	25.29 ± 2.36	0.041 ± 0.011
	10-20%	0.054 ± 0.015	0.923 ± 0.021	31.71 ± 1.82	0.010 ± 0.002	24.74 ± 1.99	0.043 ± 0.011
	20-35%	0.077 ± 0.017	0.890 ± 0.025	29.92 ± 1.72	0.014 ± 0.003	24.72 ± 1.92	0.043 ± 0.010
	ALL	0.031 ± 0.025	0.957 ± 0.036	36.85 ± 5.39	0.006 ± 0.004	25.62 ± 2.72	0.040 ± 0.013

Figure: Inpainting performance analysis.

Results II

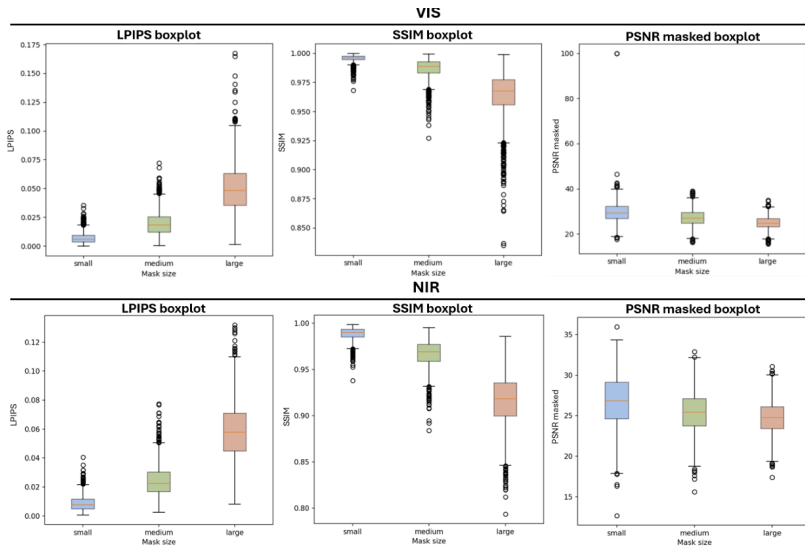


Figure: Boxplots between mask ratios and performance metrics.

Comparison to Related Work

Study	Resolution	Dataset	Spectrum	Mask shape	Mask size/ratio	No. images	PSNR	PSNR overall	SSIM	SSIM overall
[ZCGZ21]	256 × 256	IITD	NIR	Square	128 × 128 64 × 64 32 × 32	448 448 448	20.220 23.015 26.221	23.152	0.670 0.749 0.775	0.731
[CZX+23a]	256 × 256	IITD	NIR	Square	64 × 64 32 × 32	448 448	29.210 26.974	28.092	— —	—
		CASIA-IrisV4-Interval	NIR	Square	64 × 64 32 × 32	597 597	26.573 26.108	26.341	— —	—
		ND-IRIS-0405	NIR	Square	64 × 64 32 × 32	2880 2880	33.197 29.881	31.539	— —	—
[CZX+23b]	256 × 256	IITD	NIR	Eyelashes/Eyelids	—	2240	22.201	22.201	—	—
		CASIA-IrisV4-Interval	NIR	Eyelashes/Eyelids	—	2639	21.857	21.857	—	—
		ND-IRIS-0405	NIR	Eyelashes/Eyelids	—	10080	24.402	24.402	—	—
[CXC+24]	256 × 256	IITD	NIR	Irregular	0–40% 10–20% 20–30% 30–40%	448 448 448 448	28.827 28.583 26.802 25.102	27.329	0.869 0.861 0.825 0.783	0.835
					0–40% 10–20% 20–30% 30–40%	527 527 527 527	31.425 30.809 29.071 26.662	29.492	0.910 0.907 0.883 0.855	0.889
		UBIRIS.V2	VIS	Irregular	0–40% 10–20% 20–30% 30–40%	450 450 450 450	33.872 33.883 31.662 29.246	32.166	0.935 0.943 0.926 0.907	0.928
					10–40% 40–60%	448 448	29.924 25.447	27.685	0.923 0.803	0.863
		CASIA-Iris-Interval	NIR	Irregular	10–40% 40–60%	597 597	34.265 28.538	31.401	0.950 0.870	0.910
					10–40% 40–60%	2880 2880	36.249 30.829	33.539	0.975 0.941	0.958
[LZ+24]	224 × 224	CASIA-Iris-Lamp+Thousand	NIR	Irregular	narrow wide	1008 1008	34.971 30.791	32.881	0.970 0.920	0.945
X-cDiT	512 × 512	CASIA-IrisV4-Interval	NIR	Irregular	0-5% 5-10% 10-20% 20-35%	2656 1012 1242 490	41.202 35.117 31.707 29.919	36.854	0.985 0.959 0.923 0.890	0.957
		I-SOCIAL-DB	VIS	Irregular	0-5% 5-10% 10-20% 20-35%	2809 1014 1180 158	45.086 37.138 32.731 30.040	40.239	0.994 0.981 0.961 0.941	0.982

Explainability

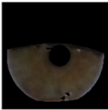
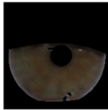
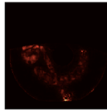
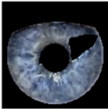
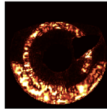
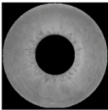
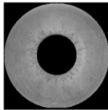
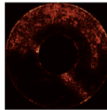
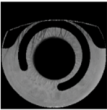
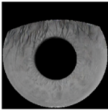
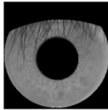
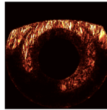
Case	Metrics	Input	Restored	Ground Truth	Error Map
C1	LPIPS:				
	SSIM:				
	PSNR:				
	MAE:				
	PSNR_masked:				
	MAE_masked:				
C2	LPIPS:				
	SSIM:				
	PSNR:				
	MAE:				
	PSNR_masked:				
	MAE_masked:				
C3	LPIPS:				
	SSIM:				
	PSNR:				
	MAE:				
	PSNR_masked:				
	MAE_masked:				
C4	LPIPS:				
	SSIM:				
	PSNR:				
	MAE:				
	PSNR_masked:				
	MAE_masked:				

Figure: Four inpainting cases selected for analysis with corresponding metrics (C1 good and C2 poor for VIS, C3 good and C4 poor for NIR).

Practical Utility

- EYERIS interactive web application
- Solid baseline for subsequent work assessing the biometric consistency of the reconstruction

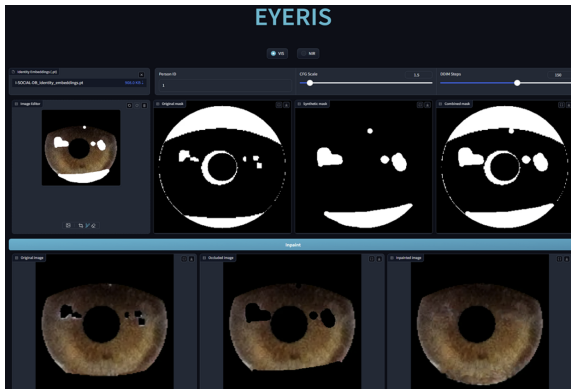


Figure: Inpainting example on a VIS image.

- Successfully implemented X-cDiT and integrated it in EYERIS software application.
- Responded to research questions:
 - **RQ1:** X-cDiT proves strong structural restoration capabilities, especially for VIS imaging.
 - **RQ2:** Overall, X-cDiT outperforms related studies, both under VIS and NIR imaging.
 - **RQ3:** Depending on iris texture complexity, occlusion dimension, location and shape, discrepancies between perceptual quality and quantitative metrics are observed.

- Assessing the performance impact of integrating X-cDiT into iris biometric recognition systems.
- Further investigating explainability by providing attention maps from the Transformer blocks.
- Experimenting with more severe occlusions.
- Exploring possible applications in genetics and ophthalmology: insights into hereditary traits, anomalies and ocular conditions.

Thank you!

Questions?

Bibliography I



CASIA.

CASIA Iris Image Database Version 4.0 (CASIA-Iris-Interval), 2010.



Ying Chen, Liang Xu, Huiling Chen, Yugang Zeng, Shubin Guo, Junkang Deng, and Ali Asghar Heidari.

Two-stage and two-discriminator generative adversarial network for the inpainting of irregularly incomplete iris images. *Displays*, 82:102626, 2024.



Ying Chen, Yugang Zeng, Liang Xu, Shubin Guo, and Huimin Gan.

Hybrid attention generative adversarial network: texture inpainting algorithm for iris defects with excellent repair performance and generalization.

Journal of Electronic Imaging, 32(3):033028, 2023.

Bibliography II



Ying Chen, Yugang Zeng, Liang Xu, Shubin Guo, Ali Asghar Heidari, Huiling Chen, and Yudong Zhang.

From coarse to fine: Two-stage deep residual attention generative adversarial network for repair of iris textures obscured by eyelids and eyelashes.

iScience, 26(7):107169, 2023.



Junkang Deng, Zhijie Chen, Ying Chen, Changle He, and Xiaodong Zhu.

High-low feature fusion generative adversarial network for the inpainting of irregularly occluded iris images.

In *Chinese Conference on Biometric Recognition*, pages 215–225. Springer, 2025.

Bibliography III



Kaiming He et al.

Deep residual learning for image recognition.

In *CVPR 2016*, pages 770–778, 2016.



R. Labati et al.

I-social-db: A labeled database of images collected from websites and social media for iris recognition.

Image and Vision Computing, 105:104058, 2021.



Chenyang Li, Zhili Zhang, et al.

I3fdm: Iris inpainting via inverse fusion of diffusion models.

In *ICASSP*, pages 1636–1640. IEEE, 2024.



William Peebles and Saining Xie.

Scalable diffusion models with transformers.

In *Proceedings of the IEEE/CVF ICCV*, pages 4195–4205, 2023.



Stability AI.

sd-vae-ft-mse, 2022.



Yugang Zeng, Ying Chen, Huimin Gan, and Zhuang Zeng.

Incomplete texture repair of iris based on generative adversarial networks.

In *Chinese Conference on Biometric Recognition*, pages 335–345. Springer, 2021.